JOURNAL OF MULTIVARIATE ANALYSIS 18, 150–168 (1986)

Approximations to the Mean Integrated Squared Error with Applications to Optimal Bandwidth Selection for Nonparametric Regression Function Estimators

WOLFGANG HÄRDLE*

Universitat Heidelberg,
Sonderforschungsbereich 123, Im Neuenheimer Feld 293,
D-6900Heidelberg 1, West Germany, and
University of North Carolina, Department of Statistics,
321 Phillips Hall 039A, Chapel Hill, North Carolina 27514

Communicated by G. Kallianpur

Discrete versions of the mean integrated squared error (MISE) provide stochastic measures of accuracy to compare different estimators of regression fuctions. These measures of accuracy have been used in Monte Carlo trials and have been employed for the optimal bandwidth selection for kernel regression function estimators, as shown in Härdle and Marron (1983), *Optimal Bandwidth Selection in Nonparametric Regression Function Estimation.* Inst. of Statistics Mimeo Series No. 1530, Univ. of North Carolina, Chapel Hill). In the present paper it is shown that these stochastic measures of accuracy converge to a weighted version of the MISE of kernel regression function estimators, extending a result of Hall (1982, *Biometrika* **69**, 383–390) and Marron (1983, *J. Multivariate Anal.* **18**, No. 2) to regression function estimation.

1. INTRODUCTION AND BACKGROUND

Let (X_1, Y_1), (X_2, Y_2),..., be independent random vectors distributed as (X, Y) with common joint probability density function $f(x, y)$ and let $m(x) = E(Y \mid X = x) = \int yf(x, y)\, dy/f_X(x)$, f_X the marginal density of X, be

* Research partially supported by the "Deutsche Forschungsgemeinschaft" SFB123, "Stochastische Mathematische Modelle," partially supported by the Air Force of Scientific Research Contract AFOSR-F49620 82 c 0009.

Received October 21, 1983; revised November 29, 1983.

AMS 1980 subject classifications: Primary 60F05; Secondary 62G05.

Keywords and phrases: stochastic measure of accuracy, nonparametric regression function estimation, optimal bandwidth selection, limit theorems, mean square error.

0047-259X/86 $3.00

the regression curve of Y on X. Let $m_n^*(x)$ denote the nonparametric kernel estimate of $m(x)$, as introduced by Nadaraya [12] and Watson [21],

$$m_n^*(x) = \hat{m}_n(x)/f_n(x) \tag{1.1}$$

where

$$\hat{m}_n(x) = n^{-1}h^{-1} \sum_{i=1}^{n} K((x - X_i)/h)\, Y_i$$

and

$$f_n(x) = n^{-1}h^{-1} \sum_{i=1}^{n} K((x - X_i)/h).$$

Here K is a kernel function and $h = h(n)$ is a sequence of "bandwidths" converging to zero as n tends to infinity.

This estimator was studied by Rosenblatt [15] who derived bias, variance, and asymptotic normality; Schuster [17] demonstrated multivariate normality at a finite number of distinct points. For further results we refer to the bibliography of Collomb [3].

In the present paper we show that

$$A_n^*(h) = n^{-1} \sum_{j \in \mathcal{J}} [m_n^*(X_j) - m(X_j)]^2, \qquad \mathcal{J} = \{j: X_j \in [0, 1]\}, \tag{1.2}$$

a stochastic measure of accuracy on the interval $[0, 1]$ for the estimate m_n^*, exhibits the same limiting behaviour as the deterministic measure

$$\text{MISE} = \int_0^1 \text{MSE}(t) f_X(t)\, dt \tag{1.3}$$

where $\text{MSE}(t)$ is the mean squared error (MSE) of $m_n^*(t)$. The proper definition of the MSE for m_n^* will be delayed to Section 2.

The result of this paper addresses two problems. First, in a survey paper, Wegman [22] was interested in comparing the mean integrated squared error (MISE) of several different density estimators. As Wegman pointed out, the computation of the actual MISE can be quite tedious. Hence, Wegman used an empirical measure of accuracy of the structure as in formula (1.2) and gave some heuristic justification. Now, since the bias/variance decomposition of regression function estimators is rather similar to that of density estimators [15, 16] it may be argued that Wegman's heuristics hold also in the regression function estimation setting. The answer is positive: It is shown here that, as $n \to \infty$, uniformly over an interval $[\underline{h}\, \bar{h}]$,

$$A_n^*(h) = \text{MISE} + o_p(\text{MISE}), \qquad h \in [\underline{h}\, \bar{h}]. \tag{1.4}$$

The appealing feature of this approximation is, that it holds uniformly in $h \in [\underline{h}\, \bar{h}]$. A Monte Carlo trial comparing different estimators of $m(x)$ (w.r.t. MISE) at different sequences of bandwidths can thus be based on $A_n^*(h)$ which is faster to compute than MISE as defined in (1.3).

Second, the approximation (1.4) contributes to the solution of the "optimal bandwidth selection" problem. As the optimal bandwidth h^* we understand that sequence $h = h(n)$ which minimizes the MISE for each n. Härdle and Marron [5] demonstrated by a crossvalidation argument that minimization (with respect to h) of $A_n^*(h)$ is asymptotically equivalent to minimization of

$$n^{-1} \sum_{j \in \mathcal{I}} [Y_j - m_n^{*(j)}(X_j)]^2, \tag{1.5}$$

where

$$m_n^{*(j)}(x) = n^{-1} h^{-1} \sum_{i \neq j} K((x - X_i)/h)\, Y_i / f_n(x)$$

is the "leave-one-out" estimator. So the result of this paper, as stated in (1.4), ensures that the minimization of (1.5) with respect to h yields the (MISE)-optimal sequence of bandwidth h^* and solves, as is shown in Härdle and Marron, a problem raised by Stone [19, Qestion 3, p. 1054].

We will not only analyze $m_n^*(x)$, as defined in (1.1), but also

$$\hat{m}_n(x)/f_X(x) \tag{1.6}$$

where f_X denotes the marginal density of X. This estimator of $m(x)$ is reasonable if we know the marginal density and is somewhat more tractable than m_n^*. The estimator (1.6) was sudied by Johnston [8], who also observed that $\hat{m}_n/f_X$ has in general a higher asymptotic variance than m_n^*.

The stochastic measure of accuracy (1.2) was defined only on the interval $[0, 1]$. It will later be assumed that the support of f_X properly contains this interval. This is due to "boundary effects," more precisely, the bias at the endpoints of the support of f_X inflates and has a slower rate than in the interior [4, 13]. Thus, defining the MISE over the whole support of f_X, would ultimately lead to the unappealing situation that the optimal bandwidth with respect to MISE would be determined in such a way that it minimizes the mean square error at the boundaries, since that is of lower order. The estimate in the interior would thus exhibit suboptimal behaviour.

The results of this paper are improvements over some previous work for several reasons. First, we do not need such strong smoothness assumptions on f_X as in Hall [6], who proves similar results in the density estimation setting. Second, our assumptions on the variance curve $V^2(t) =$

$\text{var}(Y|X=t)$ and the range of allowable bandwidths are considerably weaker than those in Johnston [8] who demonstrates a Gaussian approximation to $(nh)^{1/2}\,[\hat{m}_n - E\hat{m}_n]$ along the same lines as Bickel and Rosenblatt [1]. Third, our work extends the result of Wong [23] who deals only with the fixed design case, i.e., X_i are nonrandom. Finally, we may note that Hall's proof would simplify if one uses the approximation provided by the Bickel and Rosenblatt paper and the outline of the proof given here for regression function estimators.

Note that although only the two-dimensional case is considered here, the proof can probably be extended to the higher dimensional case where we observe a $(d+1)$-dimensional random vector $(X_1,\ldots, X_d, Y)$, $d>1$. The assumptions will be different in that case, since it is still unknown whether the multivariate empirical process can be strongly approximated by Brownian bridges with rates comparable to those in the univariate or bivariate case. This approximation technique by Brownian bridges, an carried out in the Appendix, is vital to our results. A similar technique, exploiting the idea of invariance principles in nonparametric regression, was used by Mack and Silverman [9] who showed weak and strong uniform consistency (in sup-norm) of m_n^*.

The outline of the paper is organized as follows. First, we prove that $\hat{m}_n(t) - E\hat{m}_n(t)$ can be uniformly (in t and h) approximated by a Gaussian process similar to that occurring in Bickel and Rosenblatt [1, p. 1974, formula (2.5)]. Second, we plug this approximating process into the formula (1.2), which defined the discrete version of MISE, and by evaluation of covariances and higher moments be finally arrive at the deterministic measure (1.3).

2. Results

We will make use of the following definition.

Definition. A function w is called Lipschitz-continuous of order α (LC(α)) iff with a constant L_w,

$$|w(t)-w(t')| \leqslant L_w\,|t-t'|^{\alpha}, \qquad 0<\alpha\leqslant 1.$$

The following assumptions fix the range of allowable bandwidths $[\underline{h}\,\bar{h}]$, determine the kernel function K and describe some smoothness of $m(t)$, $\text{var}(Y|X=t)$, and $f_X(t)$:

(A1) Let $\{\underline{h}_n\}$ denote a sequence for which there is an $\varepsilon>0$ so that

$$\lim_{n\to\infty} \underline{h}_n n^{1/3-\varepsilon}/\log n = 0, \qquad \lim_{n\to\infty} \underline{h}_n n^{1/2-\varepsilon} = \infty$$

and let $\{\bar{h}_n\}$ denote a sequence for which

$$\lim_{n\to\infty} \bar{h}_n = 0, \qquad \lim_{n\to\infty} \bar{h}_n \log n = \infty.$$

Assume from $h = h(n)$ that it satisfies

$$\underline{h} \leqslant h \leqslant \bar{h}.$$

(A2) There exists a sequence of positive constants $\{a_n\} \uparrow \infty$ and a $c < \infty$ such that

$$\sup_{\underline{h}\leqslant h\leqslant \bar{h}} h^{-3} \int_{|y|>a_n} y^2 f_Y(y)\, dy \leqslant c, \qquad f_Y \text{ the marginal density of } Y$$

$$\lim_{n\to\infty} \sup_{0\leqslant x\leqslant 1} \int_{|y|>a_n} y^2 f(x, y)\, dy = 0$$

$$\lim_{n\to\infty} \sup_{\underline{h}\leqslant h\leqslant \bar{h}} n^{-1/2} h^{-1/2} a_n (\log n)^2 = 0$$

$$|g_n(x)| = \left| \int_{-a_n}^{a_n} y^2 f(x, y)\, dy \right| \geqslant \eta > 0 \qquad \text{for all } 0 \leqslant x \leqslant 1,\ n \geqslant 1.$$

$$\int |d_u[g_n(uh)]| = o(\{\log(1/h)\}^{1/2}).$$

(A3) The functions $S^2(t) = E[Y^2 \mid X = t]$, $f_X(t)$ and $m(t)$ are LC(α) with $\alpha > \frac{1}{2}$ and are all of bounded variation. The marginal density of X is bounded from below:

$$\inf_{0\leqslant t\leqslant 1} f_X(t) \geqslant \gamma > 0.$$

(A4) The kernel function K is differentiable with K' of bounded variation and fulfills

$$\int K(u)\, du = 1 \qquad \text{support}\{K\} \subset [-A, A].$$

K is not assumed to be positive.

By straightforward computations it can be shown that g_n is LC(α), $\alpha > \frac{1}{2}$ and of bounded variation by assumption (A3) on $S^2(t)$ and $f_X(t)$. It is also not hard to see that if g_n is LC(1) then the last condition in (A3) follows. Note that the set of assumptions in (A2) holds if Y is bounded ($a_n = \log\log n$), an assumption that is often made in other papers, to avoid conditions on moments of Y as in (A2). (A2) also holds, if $a_n = n^\beta$, β small, while (X, Y) are jointly normally distributed. For simplicity of notation, we will not explicitly write the indices of $\bar{h}$, $\underline{h}$ h.

The following results show that the approximation (1.4) holds for both $\hat{m}_n/f_X$ and m_n^*. Only the proof of Theorem 1 (dealing with $\hat{m}_n/f_X$) will be given in full detail since the result for m_n^* can be obtained quite analogously. Let us define

$$\beta_k = \int_{-A}^{A} K^2(u)\, du$$

and

$$\hat{b}_n(t) = f_X^{-1}(t) \int_{-A}^{A} K(u)[m(t-uh) f_X(t-uh) - m(t) f_X(t)]\, du,$$

the bias of $\hat{m}_n/f_X$.

THEOREM 1. *Assume that* (A1) to (A4) hold and $\hat{b}_n(t)$ is of bounded variation. Then uniformly over $h \in [\underline{h}\ \bar{h}]$

$$\begin{aligned}
\hat{A}_n(h) &= n^{-1} \sum_{j \in \mathcal{J}} [\hat{m}_n(X_j)/f_X(X_j) - m(X_j)]^2 \\
&= (nh)^{-1} \beta_k \int_0^1 S^2(t)\, dt \\
&\quad + \int_0^1 [\hat{b}_n(t)]^2 f_X(t)\, dt \\
&\quad + o_p\left((nh)^{-1} + \int_0^1 [\hat{b}_n(t)]^2\, dt \right) \\
&= \mathrm{MISE}[\hat{m}_n/f_X] + o_p(\mathrm{MISE}).
\end{aligned}$$

Assume that f_X is d_1-times continuously differentiable and m is d_2-times continuously differentiable. Then, as in Rosenblatt [16], the bias $\hat{b}_n(t)$ would read as

$$\hat{b}_n(t) \simeq h^d \Lambda_d p^{(d)}(t)/f_X(t), \qquad p = mf_X,\ d = d_1 \wedge d_2$$

provided that K satisfies $\int u^j K(u)\, du = 0$, $j = 1, \ldots, d-1$, and $\int u^d K(u)\, du = d! \Lambda_d$. Many papers in nonparametric regression function estimation assume such a kind of differentiability as above and are dealing with methods to balance the contribution from the variance and the bias (see [3] for a review).

In a similar manner define $b_n^*(t)$, the bias of $m_n^*(t)$, as follows

$$b_n^*(t) = f_X^{-1}(t) \int_{-A}^{A} K(u)[m(t-uh) - m(t)] f_X(t-uh)\, du.$$

Where the expression "bias" has to be understood as the expected value of $f_X^{-1}[\hat{m}_n - mf_n]$, $f_n(t) = n^{-1}h^{-1}\sum_{i=1}^{n} K((t-X_i)/h)$ a density estimate of the marginal density f_X. This is justified by the observation that

$$m_n^* - m = [m_n - mf_n]/f_x + o_p(\hat{m}_n - mf_n)$$

(see [5]) and that moments of m_n^* need not exist in general [15].

The next theorem shows how $A_n^*(h)$ approximates the MISE.

THEOREM 2. *Assume that* (A1) *to* (A4) *hold and that* $b_n^*(t)$ *is of bounded variation. Then uniformly over* $h \in [\underline{h}\, \bar{h}]$,

$$\begin{aligned} A_n^*(h) &= n^{-1} \sum_{j \in \mathcal{I}} [m_n^*(X_j) - m(X_j)]^2 \\ &= (nh)^{-1} \beta_k \int_0^1 V^2(t)\, dt \\ &\quad + \int_0^1 [b_n^*(t)]^2 f_X(t)\, dt \\ &\quad + o_p\left((nh)^{-1} + \int_0^1 [b_n^*(t)]^2\, dt\right) \\ &= \mathrm{MISE}[m_n^*] + o_p(\mathrm{MISE}), \end{aligned}$$

where $V^2(t) = S^2(t) - m^2(t)$.

Note that the variance terms and the bias terms of the two estimators $\hat{m}_n/f_X$ and m_n^* are completely different. Since $V^2(t) \leqslant S^2(t)$, the Nadaraya–Watson estimator $m_n^*(t)$ attains in general a smaller (asymptotic) variance than $\hat{m}_n/f_X$. This was also observed by Johnston [8]. The condition "$nh^5 \to 0$", appearing in the work of the latter, implies that the bias vanishes asymptotically faster than the variance. Therefore, any difference in bias terms does not show up in that work. It would be interesting to find a similar comparison of bias terms, but this would lead to complicated and rather unnatural assumptions on derivatives of m and f_X, as can be seen from the formula for $\hat{b}_n$, following Theorem 1.

3. THE PROOFS

We shall prove Theorem 1 in full detail, the proof of Theorem 2 will only be sketched since the technical details are similar to the proof of Theorem 1. $F(x, y)$ will denote the joint cumulative distribution function (*df*) of (X, Y) and $F_n(x, y)$ will denote the two-dimensional empirical *df*,

defined as usual. It is understood throughout these proofs that o, 0 in remainder terms are uniform over $h \in [\underline{h}\, \bar{h}]$.

Proof of Theorem 1. The basic decomposition is

$$\hat{m}_n(t)/f_X(t) - m(t) = \hat{Y}_n(t) + \hat{b}_n(t) \tag{3.1}$$

where

$$\hat{Y}_n(t) = f_X^{-1}(t)\, h^{-1} \iint_{-\infty}^{\infty} yK((t-x)/h)\, d[F_n(x, y) - F(x, y)].$$

In the Appendix it is shown that

$$\begin{aligned} Y_{o,n}(t) &= [S^2(t)/f_X(t)]^{-1/2}\, \hat{Y}_n(t) \\ &= n^{-1/2}h^{-1} \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x) + o_p(n^{-1/2}h^{-1/2}), \end{aligned}$$

where the remainder term is uniform in t. The basic decomposition (3.1) now reads

$$\hat{m}_n(t)/f_X(t) - m(t) = n^{-1/2}h^{-1/2}V_n(t) + \hat{b}_n(t) + \rho_n \tag{3.2}$$

where $\rho_n = o_p(n^{-1/2}h^{-1/2})$ is uniformly in t and

$$V_n(t) = [S^2(t)/f_X(t)]^{1/2}\, h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x). \tag{3.3}$$

Using (3.2) and (3.3) the stochastic measure of accuracy is then

$$\begin{aligned} \hat{A}_n(h) = &\int_0^1 [\hat{b}_n(t)]^2\, dF_{X,n}(t) \\ &+ n^{-1}h^{-1} \int_0^1 V_n^2(t)\, dF_{X,n}(t) \\ &+ 2n^{-1/2}h^{-1/2} \int_0^1 \hat{b}_n(t)\, V_n(t)\, dF_{X,n}(t) \\ &+ \rho_n \left\{ 2 \left[\int_0^1 \hat{b}_n(t)\, dF_{X,n}(t) + n^{-1/2}h^{-1/2} \int_0^1 V_n(t)\, dF_{X,n}(t) \right] + \rho_n \right\}, \end{aligned}$$

where $F_{X,n}$ denotes the empirical distribution function of $\{X_i\}_{i=1}^n$. This can be rewritten as

$$\hat{A}_n(h) = n^{-1} \sum_{j \in \mathcal{J}} [\hat{b}_n(X_j)]^2$$
$$+ n^{-1}h^{-1}[U_{n1} + U_{n2}]$$
$$+ 2n^{-1/2}h^{-1/2}[U_{n3} + U_{n4}]$$
$$+ \rho_n \left\{ 2 \left[\int_0^1 \hat{b}_n(t)\, dF_{X,n}(t) + n^{-1/2}h^{-1/2} \int_0^1 V_n(t)\, dF_{X,n}(t) \right] + \rho_n \right\}$$

where

$$U_{n1} = \int_0^1 V_n^2(t) f_X(t)\, dt$$

$$U_{n2} = \int_0^1 V_n^2(t)\, d[F_{X,n}(t) - F_X(t)]$$

$$U_{n3} = \int_0^1 V_n(t)\, \hat{b}_n(t) f_X(t)\, dt$$

$$U_{n4} = \int_0^1 V_n(t)\, \hat{b}_n(t)\, d[F_{X,n}(t) - F_X(t)].$$

We now show that the limits of U_{ni}, $i = 1, 2, 3, 4$ give us the desired limit behaviour of $\hat{A}_n(h)$. We may note that the approximations, as carried out in Bickel and Rosenblatt [1], would have led to a process similar to $V_n(t)$ when estimating a density. So the technique developed here, would be useful in density estimation also and would provide an alternative proof of Hall's [6] result on stochastic measures of accuracy for density estimators.

Let us begin with the limit behaviour of U_{n1}. Note first that

$$EU_{n1} = \int_0^1 E \left\{ h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x) \right\}^2 S^2(t)\, dt$$
$$= \int_0^1 h^{-1} \int_{-\infty}^{\infty} K^2((t-x)/h)\, dx\, S^2(t)\, dt$$
$$= \int_0^1 \int_{-A}^{A} K^2(u)\, S^2(t-uh)\, du\, dt$$
$$= \beta_k \int_0^1 S^2(t)\, dt + o(1).$$

where the remainder term is uniform in h, since $S^2(t)$ is LC(α), $\alpha > \frac{1}{2}$ by assumption (A3). To show that

$$U_{n1} \xrightarrow{p} \int_{-A}^{A} K^2(u)\, du \int_0^1 S^2(t)\, dt \tag{3.4}$$

we demonstrate $E(U_{n1}^2) \sim (EU_{n1})^2$. The statement (3.4) will then follow from Chebyshev's inequality.

Since $Z(t) = h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x)$ is a Gaussian process we conclude by the Isserlis [7] formula

$$\begin{aligned} EU_{n1}^2 &= \int_0^1 \int_0^1 \{EZ^2(t_1)\, EZ^2(t_2) + 2[E[Z(t_1)\, Z(t_2)]]^2\} \\ &\quad \times S^2(t_1)\, S^2(t_2)\, dt_1\, dt_2 \\ &= \int_0^1 \int_0^1 S^2(t_1)\, S^2(t_2) \\ &\quad \times \left\{ h^{-2} \int K^2((t_1 - x_1)/h)\, dx_1 \int K^2((t_2 - x_2)/h)\, dx_2 \right. \\ &\quad \left. + 2h^{-2} \left[\int K((t_1 - x)/h)\, K((t_2 - x)/h)\, dx \right]^2 \right\} dt_1\, dt_2. \end{aligned}$$

The first summand satisfies

$$\begin{aligned} &\int_0^1 \int_0^1 S^2(t_1)\, S^2(t_2)\, h^{-2} \int K^2((t_1 - x_1)/h)\, dx_1 \int K^2((t_2 - x_2)/h)\, dx_2\, dt_1\, dt_2 \\ &\quad = \left[\beta_k \int_0^1 S^2(t)\, dt \right]^2 + O(h) \end{aligned}$$

by assumption (A4) on the kernel K.

The second summand satisfies

$$\int_0^1 \int_0^1 S^2(t_1)\, S^2(t_2)\, 2h^{-2} \left[\int K((t_1 - x)/h)\, K((t_2 - x)/h)\, dx \right]^2 dt_1\, dt_2 = O(h)$$

by evaluation of the integral inside the $[\cdot]$-brackets. This shows that

$$U_{n1} = \beta_k \int_0^1 S^2(t)\, dt + o_p(1).$$

Next we show that

$$U_{n2} = O_p(n^{-1/2} h^{-1}) \tag{3.5}$$

Define $H_n(t) = F_{X,n}(t) - F_X(t)$ and $Z_n(t) = \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x)$. We obtain by partial integration,

$$hU_{n2} = -2\int_0^1 H_n(t)\, q(t)\, Z_n(t)\left[h^{-1}q(t)\int_{-\infty}^{\infty} K'((t-x)/h)\, dW(x)\right] dt$$

$$-2\int_0^1 H_n(t)\, q(t)\, Z_n^2(t)\, dq(t)$$

$$+hH_n(t)\, V_n^2(t)|_0^1,$$

where $q(t) = S^2(t)/f_X(t)$.

Now since $H_n(t) = O_p(n^{-1/2})$ uniformly in t and $V_n^2(t_0) = O_p(1)$, $t_0 = 0, 1$, as is easily verified by Chebyshev's inequality, we only have to consider the first two summands in the equality above.

These are further estimated by Schwarz's inequality, which shows that the absolute value of the sum of both is dominated by

$$n^{-1/2}\sup_{0\le t\le 1}|n^{1/2}H_n(t)| \times \left\{S_1\left[\int_0^1 [h^{-1/2}Z_n(t)]^2\, dt\right]^{1/2}\right.$$

$$\times\left[\int_0^1\left[h^{-1/2}\int_{-\infty}^{\infty} K'((t-x)/h)\, dW(x)\right]^2 dt\right]^{1/2}$$

$$\left.+S_2\sup_{0\le t\le 1}|Z_n^2(t)|\int_0^1 |dq(t)|\right\},$$

where $S_1 = \sup_{0\le t\le 1} q^2(t)$ and $S_2 = \sup_{0\le t\le 1} q(t)$.

By Chebyshev's inequality we have

$$\int_0^1\left[h^{-1/2}\int_{-\infty}^{\infty} L((t-x)/h)\, dW(x)\right]^2 dt = O_p(1)$$

where L is either K or K'. Integration by parts applied to $Z_n^2(t)$ show immediately that $\sup_{0\le t\le 1} Z_n^2(t) = O_p(1)$, therefore (3.5) holds. Now, since

$$EU_{n3}^2 = \int_0^1\int_0^1\left\{h^{-1}\int_{-\infty}^{\infty} K((t_1-x)/h)\, K((t_2-x)/h)\, dx\right\}$$

$$\times \hat{b}_n(t_1)\,\hat{b}_n(t_2)\, q(t_1)\, q(t_2)\, dt_1\, dt_2$$

$$\le o\left(\int_0^1 [\hat{b}_n(t)]^2\, dt\right)$$

by an application of Schwarz's inequality, we conclude that

$$U_{n3} = o_p\left(\left[\int_0^1 [\hat{b}_n(t)]^2\, dt\right]^{1/2}\right). \tag{3.6}$$

The term U_{n4} is estimated again by a partial integration argument as follows,

$$U_{n4} = h^{-1/2} \int_0^1 H_n(t)\, \hat{b}_n(t)\, h^{-1} q(t) \int_{-\infty}^{\infty} K'((t-x)/h)\, dW(x)\, dt$$

$$+ h^{-1/2} \int_0^1 H_n(t)\, \hat{b}_n(t)\, Z_n(t)\, dq(t)$$

$$+ h^{-1/2} \int_0^1 H_n(t)\, q(t)\, Z_n(t)\, d\hat{b}_n(t)$$

$$+ H_n(t)\, V_n(t)\, \hat{b}_n(t)|_0^1 = T_{1n} + T_{2n} + T_{3n} + T_{4n},$$

where, as for the computations for U_{n2}, $H_n(t) = F_{X,n}(t) - F_X(t)$, and $Z_n(t) = \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x)$. The last summand T_{4n} is obviously $O_p(n^{-1/2}) = o_p(n^{-1}h^{-1})$ by (A1).

The first term, T_{1n}, can be estimated as follows:

$$|T_{1n}| \leqslant n^{-1/2}h^{-1} \sup_{0 \leqslant t \leqslant 1} |n^{1/2} H_n(t)| \left[\int_0^1 [\hat{b}_n(t)]^2\, dt \right]^{1/2}$$

$$\times S_2 \left[\int_0^1 \left[h^{-1/2} \int_{-\infty}^{\infty} K'((t-x)/h)\, dW(x) \right]^2 dt \right]^1.$$

Now, since $\int_0^1 [h^{-1/2} \int_{-\infty}^{\infty} K'((t-x)/h)\, dW(x)]^2\ dt = O_p(1)$ and $n^{1/2} \sup_{0 \leqslant t \leqslant 1} |H_n(t)| = O_p(1)$, we conclude that

$$T_{1n} = O_p\left(n^{-1/2}h^{-1} \left[\int_0^1 [\hat{b}_n(t)]^2\, dt \right]^{1/2} \right).$$

The terms T_{2n} and T_{3n} are estimated in a similar fashion as we did estimate the terms of U_{n2} employing the Lipschitz continuity of $\hat{b}_n(t)$ and $q(t)$ and we thus obtain

$$T_{2n} = O_p(n^{-1/2}) = o_p(n^{-1}h^{-1}),$$
$$T_{3n} = O_p(n^{-1/2}) = o_p(n^{-1}h^{-1}).$$

This shows finally that

$$U_{n4} = O_p\left(n^{-1/2}h^{-1} \left[\int_0^1 [\hat{b}_n(t)]^2\, dt \right]^{1/2} \right) + o_p(n^{-1}h^{-1}). \tag{3.8}$$

It remains to show that

$$\int_0^1 [\hat{b}_n(t)]^2\, d[H_n(t)] = O_p(n^{-1/2}) = o_p(n^{-1}h^{-1}). \tag{3.9}$$

Again by partial integration we have that the LHS of (3.9) is

$$-2\int_0^1 H_n(t)\,\hat{b}_n(t)\,d\hat{b}_n(t) + H_n(t)\,\hat{b}_n^2(t)\Big|_0^1.$$

As before the last summand is $O_p(n^{-1/2})$ and so is the first summand. Now, putting together (3.5) to (3.9) we finally have that

$$\hat{A}_n(h) = \int_0^1 [\hat{b}_n(t)]^2 f_X(t)\,dt + n^{-1}h^{-1}\beta_k \int_0^1 S^2(t)\,dt$$
$$+ o_p\left(n^{-1}h^{-1} + \int_0^1 [\hat{b}_n(t)]^2\,dt\right)$$

which proves the theorem.

Proof of Theorem 2. This proof goes mainly along the lines of the proof of Theorem 1. From Härdle and Marron [5, formula (2.4)], we have

$$m_n^*(t) - m(t) = Y_n^*(t) + b_n^*(t) + o_p\left(n^{-1/2}h^{-1/2} + \int_0^1 [b_n^*(t)]^2\,dt\right) \tag{3.10}$$

where

$$b_n^*(t) = f_X^{-1}(t)h^{-1}\int_{-\infty}^{\infty} K((t-u)/h)[m(u)-m(t)]f_X(u)\,du$$

and

$$Y_n^*(t) = f_X^{-1}(t)\,h^{-1}\iint_{-\infty}^{\infty} [y-m(t)]\,K((t-x)/h)\,d[F_n(x,y)-F(x,y)].$$

This process can now be approximated as $\hat{Y}_n(t)$ (see the Appendix) but with $V^2(t) = S^2(t) - m^2(t)$ in the place of $S^2(t)$. So we obtain that

$$Y_{o,n}^*(t) = [V^2(t)/f_X(t)]^{-1/2}\,Y_n^*(t)$$
$$= n^{-1/2}h^{-1}\int_{-\infty}^{\infty} K((t-x)/h)\,dW(x) + o_p(n^{-1/2}h^{-1/2})$$

uniformly in t. The decomposition (3.10) then reads as

$$m_n^*(t) - m(t) = b_n^*(t) + n^{-1/2}h^{-1/2}V_n^*(t) + \rho_n^* \tag{3.11}$$

where

$$\rho_n^* = o_p\left(n^{-1/2}h^{-1/2} + \int_0^1 [b_n^*(t)]^2\,dt\right)$$

and

$$V_n^*(t) = [V^2(t)/f_X(t)]^{1/2} h^{-1/2} \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x).$$

We then carry out the same procedures as for $V_n(t)$ in the proof of Theorem 1.

APPENDIX

It is shown here that the variance terms in (3.1) can be approximated by a sequence of Gaussian processes. The crucial step in these approximations is provided by the following lemma, due to Tusnàdy [20].

LEMMA 1. *Let* $T(x, y) = (F_X, F_{Y|X})(x, y)$ *be the Rosenblatt transformation* [14]. *Then on a suitable probability space there exists a sequence of Brownian bridges* $B_n(x', y')$ *on* $[0,1] \times [0, 1]$ *such that*

$$\sup_{x,y} |[F_n(x, y) - F(x, y)] - n^{-1/2} B_n(T(x, y))| = O_p(n^{-1}[\log n]^2).$$

It is next shown that $\hat{Y}_n(t)$ *can be approximated (uniformly in* t*) by Gaussian processes. For this define*

$$Y_{0,n}(t) = [S^2(t)/f_X(t)]^{-1/2}\, \hat{Y}_n(t)$$

$$Y_{1,n}(t) = [S^2(t) f_X(t)]^{-1/2} h^{-1} \iint_{\Gamma_n} yK((t-x)/h)\, d[F_n(x, y) - F(x, y)]$$

where $\Gamma_n = \{|\, y| \leqslant a_n\}$,

$$Y_{2,n}(t) = [S_n^2(t)/S^2(t)]^{-1/2}\, Y_{1,n}(t)$$

where $S_n^2(t) = E[\, Y^2 I(|\, y| \leqslant a_n)|\, X = t]$,

$$Y_{3,n}(t) = [S_n^2(t) f_X(t)]^{-1/2} h^{-1} n^{-1/2} \iint_{\Gamma_n} yK((t-x)/h)\, dB_n(T(x, y))$$

where $\{B_n\}$ *is the sequence of Brownian bridges as in Lemma* 1.

$$Y_{4,n}(t) = [S_n^2(t) f_X(t)]^{-1/2} h^{-1} n^{-1/2} \iint_{\Gamma_n} yK((t-x)/h)\, dW_n(T(x, y))$$

where $\{W_n\}$ is a sequence of Wiener processes used in constructing $\{B_n\}$ as

$$B_n(x', y') = W_n(x', y') - x'y'W_n(1, 1) \qquad [20]$$

$$Y_{5,n}(t) = [S_n^2(t) f_X(t)]^{-1/2} h^{-1} n^{-1/2} \times \int_{-\infty}^{\infty} [S_n^2(x) f_X(x)]^{1/2} K((t-x)/h)\, dW(x)$$

$$Y_{6,n}(t) = n^{-1/2} h^{-1} \int_{-\infty}^{\infty} K((t-x)/h)\, dW(x)$$

where $W(x)$ is a standard Wiener process on $(-\infty, \infty)$.

For the following lemmas $\| Y \|$ will denote $\sup_{0 \leqslant t \leqslant 1} | Y(t) |$.

LEMMA 2. $\| Y_{0,n} - Y_{1,n} \| = o_p(n^{-1/2} h^{-1/2})$.

Proof. We have to show that $\| U_n \| \to^p 0$, where

$$U_n(t) = n^{1/2} h^{-1/2} \iint_{|Y| > a_n} yK((t-x)/h)\, d[F_n(x, y) - F(x, y)]$$

$$= \sum_{i=1}^{n} X_{n,i}(t)$$

and

$$X_{n,i}(t) = (nh)^{-1/2} \{ Y_i K((t - X_i)/h) \cdot I(|Y_i| > a_n) - E[Y \cdot I(|Y| > a_n) K((t-X)/h)] \}.$$

Note that $EX_{n,i}(t) = 0$ for all t and that $X_{n,i}(\cdot)$ are independent, identically distributed for each n. Therefore

$$EX_{n,i}^2(t) \leqslant n^{-1} h^{-1} \sup |K|^2 \int_{|y| > a_n} y^2 f_Y(y)\, dy \qquad (4.2)$$

establishes $U_n(t) \to^p 0$ for each t by assumption (A2). By (A4) and the Cauchy–Schwarz inequality we have

$$E\, |U_n(t) - U_n(t_1)|\, |U_n(t_2) - U_n(t)| \leqslant M_0 h^{-3} |t_1 - t|\, |t_2 + t| \int_{|y| > a_n} y^2 f_Y(y)\, dy,$$

establishing by (A2) tightness of $U_n(t)$ [2, Theorem 15.6]. ∎

Note that the proof of this lemma was done as in Johnston's paper, but note also that our assumption is somewhat weaker than his, since we are employing Lemma 1, due to Tusnàdy [20], establishing a faster rate for the two-dimensional empirical process.

LEMMA 3. $\| Y_{1,n} - Y_{2,n} \| = o_p(n^{-1/2}h^{-1/2})$.

Proof. Define $g(t) = S^2(t) f_X(t)$, $g_n(t) = S_n^2(t) f_X(t)$. We must show that

$$\sup_{0 \leq t \leq 1} \left\{ | g(t)^{-1/2} - g_n(t)^{-1/2} | \cdot \left| h^{-1} \iint_{\Gamma_n} yK((t-x)/h)\, d[F_n(x,y) - F(x,y)] \right| \right\} = o_p(n^{-1/2}h^{-1/2}).$$

Now, from Johnston [8] we have that the second factor inside the curly brackets is $O_p(n^{-1/2}h^{-1/2})$ and from the mean value theorem

$$| g_n^{-1/2} - g^{-1/2} | = | g_n - g | \cdot | \tfrac{1}{2}\xi_n^{-3/2} |,$$

where ξ_n is between g_n and g. Since g_n, g are bounded away from zero by assumption (A3), $\| \xi_n^{-3/2} \|$ is a bounded sequence. Finally, from (A2) it follows that $\| g_n - g \| \to 0$ and thus the lemma follows.

LEMMA 4. $\| Y_{2,n} - Y_{3,n} \| = o_p(n^{-1/2}h^{-1/2})$.

Proof. Using integration by parts (see [8, Lemma A.5] for details), we obtain

$$n^{1/2}h^{1/2} | g_n(t) |^{1/2} | Y_{2,n}(t) - Y_{3,n}(t) |$$
$$= O_p(n^{-1/2}(\log n)^2)\, h^{-1/2} \left\{ 4a_n \int_{-A}^{A} | K'(u) |\, du + 4a_n [| K(A) | + | K(-A) |] \right\}$$
$$= O_p(n^{-1/2}h^{-1/2}a_n(\log n)^2)$$

uniformly in t. The proof thus follows using assumption (A2).

LEMMA 5. $\| Y_{3,n} - Y_{4,n} \| = o_p(n^{-1/2}h^{-1/2})$.

Proof. Since the Jacobian of the transformation T, introduced in Lemma 1, is $f(x, y)$, we have by Masani [11, Theorem 5.19],

$$n^{1/2} | Y_{3,n}(t) - Y_{4,n}(t) |$$
$$\leq | g_n(t)^{-1/2}\, h^{-1} \iint_{\Gamma_n} yK((t-x)/h) f(x,y)\, dx\, dy | \cdot | W_n(1,1) |.$$

So we finally have

$$n^{1/2} \| Y_{3,n} - Y_{4,n} \| \leqslant | W_n(1, 1) | \lambda_1 h^{-1} \int | K((t-x)/h) | \, dx$$

where λ_1 is a constant ($\lambda_1 = \sup_{0 \leqslant t \leqslant 1} | m(t) f_X(t) |$). This proves the lemma. Note that $Y_{4,n}(t)$ is a zero mean Gaussian process with covariance

$$\begin{aligned}
&\operatorname{cov}\{ Y_{4,n}(t_1), Y_{4,n}(t_2) \} \\
&\quad = [S_n^2(t_1) f_X(t_1)]^{-1/2} [S_n^2(t_2) f_X(t_2)]^{-1/2} \\
&\qquad \times n^{-1} h^{-2} \iint_{\Gamma_n} y^2 K((t_1 - x)/h) \, K((t_2 - x)/h) f(x, y) \, dx \, dy \\
&\quad = \operatorname{cov}\{ Y_{5,n}(t_1), Y_{5,n}(t_2) \}.
\end{aligned}$$

So both $Y_{4,n}$ and $Y_{5,n}$ are Gaussian processes with the same covariance structure and can thus be identified.

LEMMA 6. $\| Y_{5,n} - Y_{6,n} \| = o_p(n^{-1/2} h^{-1/2})$.

Proof. Note that by assumption (A3) on $g_n(t) = S_n^2(t) f_X(t)$,

$$G_{n,t}(u) = [g_n(t)]^{-1/2} \{ [g_n(t - uh)]^{1/2} - [g_n(t)]^{1/2} \}$$

is also LC(α), $\alpha > \frac{1}{2}$, i.e.,

$$| G_{n,t}(u) - G_{n,t}(u') | \leqslant L_G h^{\alpha} | u - u' |^{\alpha}, \qquad \alpha > \tfrac{1}{2},$$

where L_G is independent of t by (A3).

The difference of interest is now

$$\begin{aligned}
&(nh)^{1/2} | Y_{5,n}(t) - Y_{6,n}(t) | \\
&\quad = h^{-1/2} \left| \int \{ [g_n(x)/g_n(t)]^{1/2} - 1 \} \, K((t-x)/h) \, dW(x) \right| \\
&\quad = | R_n(t) |.
\end{aligned}$$

We will now show that $\sup_{0 \leqslant t \leqslant 1} | R_n(t) | = o_p(1)$. By partial integration we have for all n and t,

$$\begin{aligned}
| R_n(t) | &\leqslant \left| h^{-1/2} \int_{-A}^{A} W(t - uh) \, G_{n,t}(u) \, K'(u) \, du \right| \\
&\quad + \left| h^{-1/2} \int_{-A}^{A} [W(t - uh) - W(t)] \, K(u) \, d[G_{n,t}(u)] \right| \\
&\quad + \left| h^{-1/2} \int_{-A}^{A} W(t) \, G_{n,t}(u) \, K'(u) \, du \right| + O_p(h^{1/2}). \\
&= R_{1,n}(t) + R_{2,n}(t) + R_{3,n}(t) + R_{4,n},
\end{aligned}$$

where $R_{4,n}$ is independent of t. The term $R_{1,n}(t)$ is estimated as in Johnston [8, Lemma 4.6, p. 411] to obtain

$$\sup_{0 \leqslant t \leqslant 1} |R_{1,n}(t)| = o_p(1).$$

We now show that

$$\sup_{0 \leqslant t \leqslant 1} |R_{2,n}(t)| = o_p(1).$$

Let $w_o(s)$ denote the modulus of continuity of $W(t)$ and let $\bar{K} = \sup_{-A \leqslant u \leqslant \leqslant A} |K(u)|$, we then have with Silberman [18, formula (7), (8), and his definitions of p, q, B],

$$|R_{2,n}(t)| \leqslant h^{-1/2}\bar{K} \int w_0(|u|h)|\, dG_{n,t}(u)|$$

$$\leqslant h^{-1/2} 16 \bar{K}\, 2^{1/2} \int_{-A}^{A} q(|u|h)\, dG_{n,t}(u)|$$

$$+ h^{-1/2} 16 \bar{K} (\log B)^{1/2} \int_{-A}^{A} p(|u|h)|\, dG_{n,t}(u)|.$$

Now following the proof of Silverman [18, Proposition 4] we see that the both summands are by assumption (A3) on $|dg_n(u)|$ of the order $o_p(1)$ uniformly in t. It remains to show that $\sup_{0 \leqslant t \leqslant 1} |R_{3,n}(t)| = o_p(1)$. This follows again from assumption (A3) on the LC(α), $\alpha > \frac{1}{2}$ condition $g_n(\cdot)$, and the following inequality:

$$\sup_{0 \leqslant t \leqslant 1} |R_{3,n}(t)| \neq \eta^{-2} \sup_{0 \leqslant t \leqslant 1} |W(t)|\, h^{-1/2} L_G h^{\alpha} \int_{-A}^{A} |u|^{\alpha} |K'(u)|\, du = o_p(1).$$

ACKNOWLEDGMENT

I am grateful to Steve Marron for helpful discussions. Ray Carroll contributed much to the approximations of the Appendix.

REFERENCES

[1] BICKEL, P., AND ROSENBLATT, M. (1973). On some global measures of the deviation of density function estimators. *Ann. Statist.* **1** 1071–1095.

[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[3] COLLOMB, G. (1981). Estimation non-parametrique de la regréssion: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.
[4] GASSER, T. AND MÜLLER, G. H. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (T. Gasser and M. Rosenblatt, Ed.), Lecture Notes in Mathematics Vol. 757, Springer-Verlag Heidelberg.
[5] HÄRDLE, W., AND MARRON, S. (1983). *Optimal Bandwidth Selection in Nonparametric Regression Function Estimation.* Institute of Statistics Mimeo Series No. 1530, University of North Carolina, Chapel Hill.
[6] HALL, P. (1982). Cross-validation in density estimation. *Biometrika* **69** 383–390.
[7] ISSERLIS, L. (1918). On a formula for the product moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* **12** 134–139.
[8] JOHNSTON, G. (1982). Probabilities of maximal deviations of nonparametric regression function estimation. *J. Multivariate Anal.* **12** 402–414.
[9] MACK, Y. P., AND SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415.
[10] MARRON, J. S. (1986). Convergence properties of an empirical error criterion for multivariate density estimation. *J. Multivariate Anal.* **18**, No. 2.
[11] MASANI, P. (1968). Orthogonally scattered measures. *Adv. in Math.* **2** 61–117.
[12] NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.
[13] RICE, T., AND ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.
[14] ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23** 470–472.
[15] ROSENBLATT, M. (1969). Conditional probability density and regression estimation. In *Multivariate Analysis* II (P. R. Krishnaiah, Ed.), pp. 25–31. Academic Press, New York.
[16] ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Stat.*, **42**, 1815–1842.
[17] SCHUSTER, E. F. (1972), Joint asymptotic distribution of the estimated regression function at a finite number of district points. *Ann. Math. Stat.*, **43**, 84–88.
[18] SILVERMAN, B. (1982). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Stat.*, **6**, 177–184.
[19] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.*, **10**, 1040–1053.
[20] TUSNÀDY, G. (1977). A remark on the approximation of the sample distribution function in the multidimensional case. *Period. Math. Hung.*, **8**, 53–55.
[21] WATSON, G. S. (1964). Smooth regression analysis. *Sankhya, Series A*, Vol. **26**, 359–372.
[22] WEGMAN, E. J. (1972). Nonparametric probability density estimation: A comparison of density estimation methods. *J. Statist. Comput. Simulation*, **1**, 225–245.
[23] WONG, W. H. (1983). On the consistency of coss-validation in kernel nonparametric regression. *Ann. Stat.*, to appear.

Printed in Belgium